Investigating DEI-related biases in GEO using Python

Hojae Lee

March 9th, 2022

Abstract

Here we investigated DEI-related biases in the Gene Expression Omnibus (GEO) for the DCMB DEI Data Challenge. We developed a quantitative DEI score on a scale of 0-10 based on sex and race provided in individual GEO samples. We analyzed 1,568 datasets out of 2,564 datasets that match the DEI Data Challenge criteria. Only 170 datasets scored above 0, indicating that nearly 90% of the datasets do not report either sex or race for their samples. The average DEI score was 3.97 out of 10, with no significant difference in scores before and after 2015. Average DEI scores vary by country, with the United States ranking 7th out of 15.

Methods

Raw DEI Dataset. Using an open source Python project called Entrez[1], GEOparse package from PyPI and Pandas, we constructed our own DEI dataset to analyze potential DEI-related biases present in the GEO database. We collected DEI-related metadata from a small subset of datasets matching the following criteria: the datasets must (1) study the organism Homo sapiens, (2) conduct experiment of type “expression profiling by array”, and (3) contain at least 100 samples per dataset. This yielded 2,564 GEO accession numbers from the GEO database as of March 9th, 2022. Of the 2,564 datasets, we analyzed 1,568 datasets (around 60%) during the challenge period between March 7th through March 9th.

The DEI dataset consists of 478,216 samples across 1,567 datasets and the following eight columns:

- `gse_id`: ID of the GEO Dataset
- `gsm_id`: ID of the Sample within GEO Series
- `contact_country`
- `submission_date`
- `sex`
- `race`
- `ethnicity`
- `age`

The dataset is available for download at https://github.com/hojaeklee/dcmb_dei_challenge.

Preprocessing / Cleaning. We preprocessed two out of four DEI-related features (sex and race) for consistency prior to computing DEI scores. For sex, raw inputs such as `[“m”, “boys”, “male”]` were replaced with “male”, while [`“f”, “girls”, “female”`] were replaced with “female”. Other values, such as 0 or 1 (e.g. ambiguous values), “refused”, etc. were mapped to NaN (i.e. not a number). We performed a similar rule-based mapping for race, e.g. `[“white”, “caucasian”, “w”, “cauc”]` mapping to “white”, and grouped into eight categories: White, Hispanic, Black, Asian, American Indian, Pacific Islander, Multiple, and Other. We noticed numerous inconsistencies and ambiguities in reporting race, which we took into consideration for computing our DEI score. We highlight a few inconsistencies or ambiguities for example:

- We noticed that “Hispanic” were underrepresented compared to other categories. This may be due to the lack of option, e.g. the US 2020 census did not consider Hispanic as race.
- Responses such as “na” or “aa” were ambiguous as to whether they stand for “North American” or “Not Available” (for “na”) or “African American” or “Asian American” (for “aa”).

We also converted the common names in the `contact_country` column to their alpha-3 codes[2] (e.g. Netherlands to NLD) according ISO 3166 international standards using the `pycountry` package. Finally, we converted the submission dates using the `datetime` package. The cleaned DEI dataset is available for download at the project’s Github page.

Computing DEI Score. We devised a quantitative DEI score per dataset based on the reported sex and race of their samples. The DEI score is a weighted sum of the following four criteria:

- Existence of “sex” characteristics
- Ratio of female to male

- Existence of “race” characteristics
- Distribution of “race” compared to that of US population

Each criterion is scored between 0 and 1, and weighted as follows for a total score between 0 (worst) and 10 (best):

$$\text{DEI Score} = 3^*\text{sex\_exist} + 2^*\text{sex\_ratio} + 4^*\text{race\_exist} + 1^*\text{race\_distribution}$$

As shown in the above equation, we put more weight on the existence of sex and race characteristics in samples, as most datasets do not report this information. We gave less weight to “race_distribution” to minimize error from inconsistencies in reporting and preprocessing errors. We now describe how each criterion were calculated in detail.

Existence of “sex” characteristics. If a sample contained either “male” or “female” (after preprocessing), the sample received 1 point, else 0 points. The overall score for a dataset is the average across its samples.

Ratio of female to male. For each dataset that reported sex, we computed a female-to-male ratio by counting each of their occurrences. The ratios ranged from 0.0 (no female/all male) up to 5.0 with no upper limit. We wanted to devise a score between 0 and 1 point, with more points awarded to datasets whose female-to-male ratio is close to 1. To do this, we used a log normal distribution of $\mu = 0 \text{ and } \sigma = 0.5$. This represents the green distribution shown in the figure below.

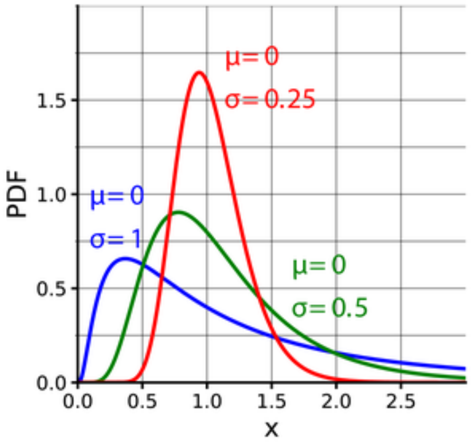

Figure. Plot of three different log-normal probability density functions. (Source: Wikipedia)

We used a log normal distribution as the values of ratios are strictly non-negative. Finally, we computed the sex ratio score as follows:

$$\text{sex\_ratio} \; = 1 - 2 * \text{abs}(0.5 \text{ - lognorm.cdf(female-to-male ratio))}$$

This yields a score between 0 and 1, and awards highest point to datasets whose female-to-male ratio is 1, and less points to those whose female-to-male ratio deviates from 1.

Existence of “race” characteristics. Identical to how we computed existence of “sex” characteristics.

Distribution of race compared to that of US population. For each dataset that reported race, we calculated a normalized Wasserstein distance of each sample’s race percentage to that of the US population. We chose the US population as the baseline as most datasets that reported race were from the USA.
Briefly, the Wasserstein distance, also known as Earth Mover’s distance, is a distance function between probability distributions ranging from 0 (identical distributions) to infinity. In our case, the race distribution represents a discrete probability distribution whose random variable can take on one of eight values: White, Hispanic, Black, Asian, Multiple, Other, Pacific Islander. We computed the Wasserstein distance against the race distribution of the US population given in the table below.

Race	Percentage
White	57.8%
Hispanic	18.7%
Black	12.1%
Asian	5.9%
Multiple	4.1%
American Indian	0.7%
Other	0.5%
Pacific Islander	0.2%

Table 1. 2020 U.S. Census, including extra category for Latino / Hispanic (Source: Wikipedia)

To normalize the values to be between 0 and 1, we divided each Wasserstein distance to a theoretical maximum Wasserstein distance. In this case, we computed the theoretical maximum Wasserstein distance as the distance between all 0 racial distribution to the racial distribution of the US population. The resulting scores award close to 1 point if the reported racial distribution per dataset is closest to that of the US population.

Results

Distribution of DEI scores: ~90% datasets have DEI score of zero. Of the 1,568 datasets analyzed, only 170 datasets (10.8%) had a DEI score of greater than zero, indicating that nearly 90% of the datasets did not report either sex or race characteristics in any of their samples. Of the 170 DEI scores calculated, the mean DEI score was 3.97 with a standard deviation of 1.19. A distribution of positive DEI scores is plotted below:

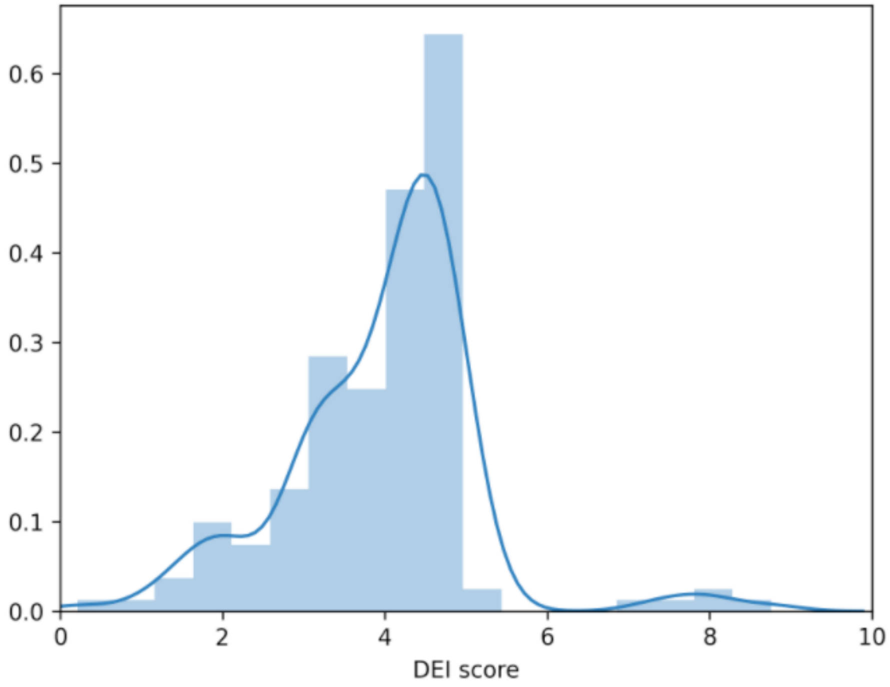

Figure 2. Distribution of DEI scores across 170 datasets. Higher DEI scores indicate less bias.

Below we report the top 5 least biased datasets and their computed DEI scores:

Dataset	DEI SCore
GSE75285	8.76
GSE60190	8.03
GSE130588	7.98
GSE71620	7.62
GSE59206	7.30

Table 2. Top 5 least biased datasets in GEO.

<u>There is no significant difference between DEI scores before and after 2015</u>. Next, we grouped and averaged the DEI scores based on their submission date into two categories: Before 2015 and After 2015. As shown in the plot below, we did not see a significant difference between distributions of DEI scores before and after 2015.

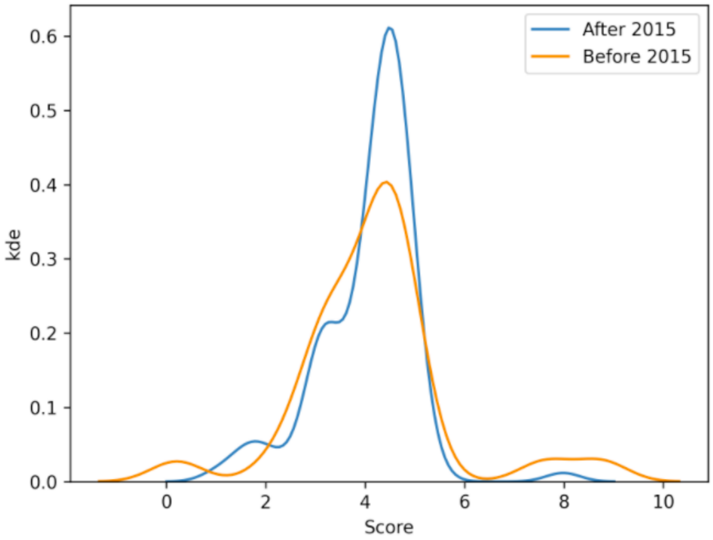

Figure 3. Distribution of DEI scores across 170 datasets. There is no significant difference between computed DEI score for datasets submitted before or after 2015.

<u>Average DEI scores across country.</u> We computed the average DEI scores for 170 datasets across 15 countries. Using `geopandas`, we plotted the average DEI score per country as shown below.

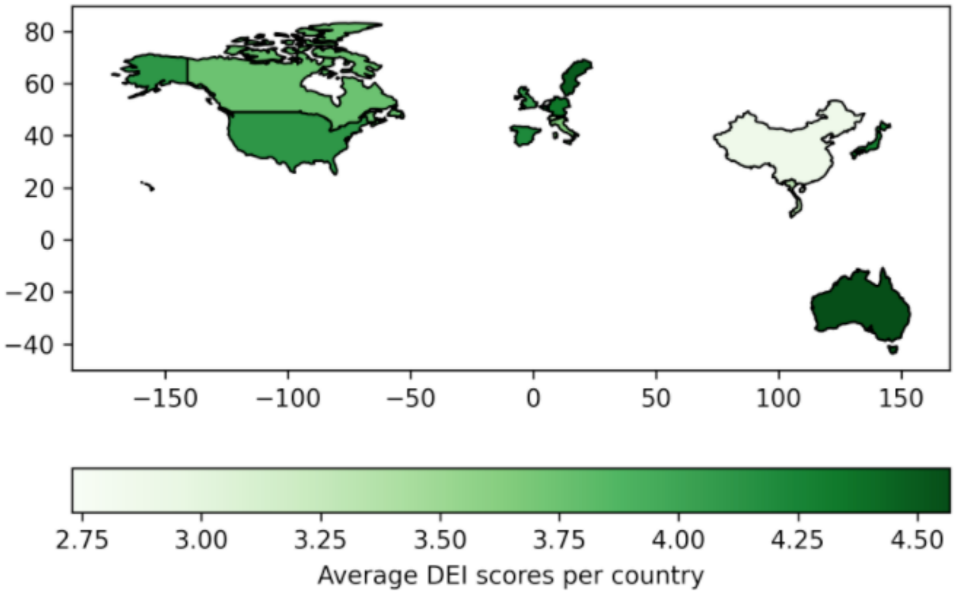

Figure 4. Average DEI scores per country.

The countries with the top five DEI scores are Australia (4.57), Belgium (4.52), Sweden (4.50), Germany (4.34), and Japan (4.28). The US ranked 7th out of 15 countries with DEI score of 4.11.

References

1. https://github.com/jordibc/entrez
2. https://www.iban.com/country-codes